MedPharmRes
journal of University of Medicine and Pharmacy at Ho Chi Minh City
homepage: http://www.medpharmres.vn/ and http://www.medpharmres.com/

Original article

Inter-Rater reliability of a professionalism OSCE developed in family medicine training University of Medicine and Pharmacy

Pham Duong Uyen Binh[a*], Pham Le An[a], Tran Diep Tuan[a], Jimmie Leppink[b]

[a]*University of Medicine and Pharmacy HCMC;*
[b]*School of Health professions education, Maastricht University, the Netherlands.*

Received January 10, 2018: Accepted April 02, 2018: Published online April 03, 2018

Abstract: A POSCE was developed and administered in 2015 to assess six professional attributes for the Family Medicine (FM) residents, University of Medicine and Pharmacy (UMP), Vietnam. This study aims at exploring inter-rater reliability in FM POSCE developed in this context when analytic rubrics were applied. **Background:** Past POSCEs showed raters' variability on applying the global marking items and holistic rating. Using analytic rubrics, unlike holistic type, will provide more rationale for assigning a certain score might influence raters' variability. Nonetheless, it is little known to what extent switching to this rubric type might influence the inter-rater reliability of POSCE. **Methods:** Before the FM professionalism module (pretest) and after this module (posttest), 36 and 42 FM residents took the POSCE respectively. The raters in the pretest included 12 teachers of FM training center. Four faculty members from different faculties were belatedly added to the post-test together with the 12 former raters. Raters' training occurred in two different times, the former took place only for the 12 FM raters before the pretest and the latter was before the posttest for the 4 belatedly-recruited. During the POSCE, one pair of raters observed all performances per station. Inter-rater reliability was measured by the differences in total scores between raters per pair using paired t-test and Pearson correlation coefficient. **Results:** In POSCE pretest, no significant difference was found between raters' scores in most pairs of raters, contrasting with that in the posttest. Most differences were noticed in the pairs of raters, in which one of the raters was the belatedly-recruited. In the pretest, moderate to strong positive correlation between raters' mean scores were found (r=0.55-0.85), similar range was seen in the post-test (r=0.47-0.87), however, the correlation slightly weakened. **Discussion and conclusion:** The FM POSCE has high inter-rater reliability on the utilization of analytic grading rubrics. An analytic rubric might help minimize the discrepancies among raters. Moreover, training raters might have been an alternative influential factor on the raters' consensus.

Key words: professionalism, OSCE, inter-rater reliability, analytic rubrics, raters' training.

1. INTRODUCTION

For the last three decades, Objective Structured Clinical Exams (OSCEs) have been used for the assessment of clinical competence, medical knowledge, interpersonal, communication skills and professionalism as part of health professions education. Despite the apparent advantages of the professionalism OSCE (POSCE) over self-answered questionnaires and work-based assessments, the psychometrics of this standardized exam has been the emerging topic in the literature. The reliability of the assessment is crucial, particularly when the aim of the POSCE is to provide the rationale for the judgment of medical novices' professionalism, as is often the case in medical school assessments [2].

Particularly, some data on reliability of POSCE that determining resident's acquisition of professionalism have been reported in several studies [4, 9]. Inter-rater reliability is one of the most concerned estimator of reliability of

*Address correspondence to this author Pham Duong Uyen Binh at University of Medicine and Pharmacy at Ho Chi Minh city, Vietnam; E-mails: binhpham2599@gmail.com
DOI: 10.32895/UMP.MPR.2.1.20

POSCE as making inferences from performance ratings requires the management of rater effects [1]. Findings from the past POSCEs showed inter-grader variability among different raters in grading same professional behaviors. Nonetheless, it is little known that the POSCE developed in Vietnam yields acceptable inter-rater reliability or less differences among raters. Therefore, the aim of this study was to investigate inter-rater reliability of the POSCE that was developed in the context of FM training in Vietnam.

2. METHOD

The POSCE

A POSCE was developed and conducted in Training Center of Family medicine, the University of Medicine and Pharmacy (UMP), Ho Chi Minh city. POSCE was administered at two different times, at the end of September, 2015 before the module of Counseling and Professionalism and at the beginning of November, 2015 in the FM orientation course.

Examiners

Only faculty raters are recruited for the POSCE. In the pretest, 12 faculty members who were teaching faculty with both an MD and MSc in the field of FM with at least 5 years of clinical practice and teaching were invited. In the posttest, 12 raters in the pretest and 4 belatedly-recruited raters from the unit of Preventive Medicine(UMP) were invited. All raters had not experienced in rating professionalism OSCE before.

Raters' training occurred in two different times, the former took place only for the 12 FM raters before the pretest and the latter was before the posttest for the 4 belatedly-recruited.

Examination procedure

All candidates rotated through six stations. In each station, FM residents interact with a Standardized Patient (SP) who portrays a scripted ailment of a specific scenario. Two raters were arranged to grade performances in a station. It was customary that during the encounter, raters completed an evaluation form that contains marking items and the 3-point rubrics which pertained to these marking items. The grading rubric comprised 3 anchors: 2-meet standard; 1-borderline; 0-below standard. Behavioral descriptors were provided in each anchor of the rubrics.

Examiners' training

Raters' training was provided before pretest and posttest consisting of four steps as follows.

Step1 and 2: Overview and Briefing

The raters viewed all scenarios and the scoring rubrics before training sessions. The author of cases and a content expert, Director of FM training center clarified any details of the cases, itemlists and the analytic rubrics.

Step 3: Familiarization with scoring criteria at each mastery level

Each group of raters participated in six one-hour training sessions for six scenarios. In each session, raters used the scoring rubrics to rate performances in three randomly-shown video clips. These clips intentionally demonstrated performances of three different mastery levels in each case, which was unknown to the raters.

Step 4: Discussion

After completing their scoring, the raters compared their scores with others' items by items. Differences in assigning score in each item to the same encounter were discussed. Differences between raters' and the expert's scores in the same video clip were also discussed. This enabled examiners to achieve consensus regarding what constituted below-standard, borderline or meet-standard performance of certain behaviors.

Statistical analysis

Descriptive statistics of the scores given by the raters were calculated using SPSS 20. Inter-rater reliability was measured by the differences in mean scores between raters using paired t-test and inter-rater agreement using Pearson correlation coefficient.

Ethical statements

Informed consent

All participants informed that their results will be analyzed for an evaluating study. They were also assured that their identities would be kept confidential. All participants gave verbal consent to join in the study. Their approvals were obtained on the exam days by having them sign on the registering paper before the exam.

3. RESULTS

Table 1 portrays the values of paired sample T-test for each pair of raters' total scores assigning to one performance in pretest and posttest. No significant difference was found between raters' mean scores in most pairs of raters in the OSCE pretest. Significant differences were found mostly in scoring the scenario of "Keeping confidentiality", "Breaking bad news", "Altruism" and "Self-awareness of limitation". However, in the OSCE posttest, differences in mean scores between raters were found in eight out of twelve pairs. Notably, raters' differences occur in all scenarios.

Table 1: The results of the paired sample T-test for raters' total scores in pretest and posttest

			Pretest			Posttest		
			T	**Df**	**Sig. (2-tailed)**	**T**	**Df**	**Sig. (2-tailed)**
Pair 1	rater1scen1 - rater2scen1	Keeping confidentiality	0.08	18	0.93	-1.31	14	Pair 1
Pair 2	rater3scen1 - rater4scen1		-8.45	23	<0.001	-2.85	15	Pair 2
Pair 3	rater1scen2 - rater2scen2	Responsibility in community	-3.41	23	<0.001	-0.93	16	Pair 3
Pair 4	rater3scen2 - rater4scen2		-1.47	18	0.15	1.38	12	Pair 4
Pair 5	rater1scen3 - rater2scen3	Disclosing medical errors	0.29	23	0.77	1.05	16	Pair 5
Pair 6	rater3scen3 - rater4scen3		-7.68	17	<0.001	1.53	15	Pair 6
Pair 7	rater1scen4 - rater2scen4	Breaking bad news	6.42	19	<0.001	-3.75	15	Pair 7
Pair 8	rater3scen4 - rater4scen4		6.52	11	<0.001	4.46	16	Pair 8
Pair 9	rater1scen5 - rater2scen5	Making altruistic decision	7.54	23	<0.001	0	14	Pair 9
Pair 10	rater3scen5 - rater4scen5		-1.84	15	0.09	4.46	16	Pair 10
Pair 11	rater1scen6 - rater2scen6	Admitting limitation	-6.28	23	<0.001	4.17	19	Pair 11
Pair 12	rater3scen6 - rater4scen6		-8.51	17	<0.001	1.68	14	Pair 12

Table 2: Paired Samples Correlations in pretest and posttest

		Pretest		**Posttest**	
		Correlation	Sig.	Correlation	Sig.
Pair 1	rater1scen1 & rater2scen1	0.69	0.004	0.65	0.002
Pair 2	rater3scen1 & rater4scen1	0.55	0.03	0.05	0.78
Pair 3	rater1scen2 & rater2scen2	0.57	0.02	0.75	0.00
Pair 4	rater3scen2 & rater4scen2	0.79	0.00	0.82	0.00
Pair 5	rater1scen3 & rater2scen3	0.78	0.00	0.87	0.00
Pair 6	rater3scen3 & rater4scen3	0.81	0.00	0.67	0.002
Pair 7	rater1scen4 & rater2scen4	0.84	0.00	0.68	0.001

		Pretest		Posttest	
Pair 8	rater3scen4 & rater4scen4	0.85	0.00	0.75	0.005
Pair 9	rater1scen5 & rater2scen5	0.81	0.00	0.59	0.003
Pair 10	rater3scen5 & rater4scen5	0.85	0.00	0.80	0.00
Pair 11	rater1scen6 & rater2scen6	0.82	0.00	0.45	0.03
Pair 12	rater3scen6 & rater4scen6	0.81	0.00	0.65	0.00

Table 2 presents the correlation between raters' total scores. Moderate to strong positive correlation between raters' mean scores were found in the pretest. Mean scores between raters in most pairs were strongly correlated, except in pair two and three where there is a positive moderate correlation between raters' mean scores. In the posttest, mean scores between raters in the other pairs were strongly correlated. However, very weak correlation was also found between raters' mean scores in pair two.

4. DISCUSSION

We found a strong consistency in grading and correlation between raters' scores for residents' performances in the POSCE. This would suggest that POSCE is able to consistently measure the candidates' professional behaviors across different raters.

The finding from this study implied the important role of analytic rubrics in achieving high consensus among raters. When using holistic rubrics, raters are believed to use their intuition to rapidly decide which category a performance falls into [6]. However, raters still analyzed what they have observed and later, applied their personal experiences to make their assignment of scores since holistic rubric provides raters with a few of what constitutes a professional behavior4. This might increase subjectiveness, thus, cause more differences among raters in evaluation of professional behaviors [4]. Therefore, the analytic rubrics comprising case-relevant marking items and behavioral descriptors that guide the raters' judgment might have lessened the raters' bias and improved the inter-rater agreement.

Lack of consensus among raters might reduce the raters' consistency in assigning scores [8]. This argument has been supported by this study. It found that most differences in total scores assigned by the pairs of raters, in which one of them was the belatedly-recruited for the POSCE posttest. Given that prior to the OSCE posttest, only these raters were involved in the training. Lack of discussion to reach consensus on how to assign scores among the former and the later raters might have caused raters' gaps despite the similar training on grading professional behaviors.

This study suggests that analytic rubrics together with several features of raters' training might improve the raters' consistency. First, practical section should be included, in which raters are exposed to candidates' samples of real performances to practice rating using the rubrics. Video clips can be an effective mean for practice if they clearly demonstrate performances at each mastery level on the rubrics. At the end of the practicing session, it is essential for all raters to compare their scores with the others' for the same encounters and open discussions on the reasons for any discrepancies in scoring [8]. This can trigger a reaching consensus process, which is valuable in bridging the gaps between raters.

Nonetheless, this is a cross-sectional study. It is impossible to conclude to what extent those abovementioned factors influenced the inter-rater reliability. Moreover, there might be other factors that affect the inter-rater reliability such as raters' professional backgrounds. Therefore, future studies should investigate multiple factors and their extent to which they affect raters' consensus in rating in POSCE. Understanding these factors helps better manage the rater effects in POSCE and other performance-based assessments of professionalism.

5. CONCLUSION

FM POSCE can is able to consistently measure the candidates' professional behaviors across different raters. Using analytic rubrics and features of raters' training which facilitates raters' practice of rating and discussion on discrepancies in scoring among raters might help improve the inter-rater reliability.

REFERENCES

1. I. Bartman, S. Smee, M. Roy. A method for identifying extreme OSCE examiners. The clinical teacher. 2013;10(1):27-31.

2. M. T. Brannick, H. T. Erol□Korkmaz, M. Prewett. A systematic review of the reliability of objective structured clinical examination scores. Medical education. 2011;45(12):1181-9.

3. W. C. Husser. Medical professionalism in the new millenium: A physician charter. Journal of the American College of Surgeons. 2003;196(1):115-8.

4. K. M. Mazor, M. L. Zanetti, E. J. Alper, D. Hatem, S. V. Barrett, V. Meterko, et al. Assessing professionalism in the context of an objective structured clinical examination: an in□depth study of the rating process. Medical education. 2007;41(4):331-40.

5. V. T. Nhan, C. Violato, P. Le An, T. N. Beran. Cross-Cultural Construct Validity Study of Professionalism of Vietnamese Medical Students. Teaching and learning in medicine. 2014;26(1):72-80.

6. M. J. Peeters. Measuring rater judgments within learning assessments—Part 2: A mixed approach to creating rubrics. Currents in Pharmacy Teaching and Learning. 2015;7(5):662-8.

7. G. Ponnamperuma, J. Ker, M. Davis, M. Medical professionalism: teaching, learning and assessment. South-East Asian Journal of Medical Education. 2012;1(1):42-8.

8. E. Schwartzman, D. I. Hsu, A. V. Law, E. P. Chung. Assessment of patient communication skills during OSCE: Examining effectiveness of a training program in minimizing inter-grader variability. Patient education and counseling. 2011;83(3):472-7.

9. Y. Y. Yang, F. Y. Lee, H. C. Hsu, W. S. Lee, C. L. Chuang, C. C. Chang, et al. Validation of the behavior and concept based assessment of professionalism competence in postgraduate first-year residents. Journal of the Chinese Medical Association. 2013;76(4):186-94.

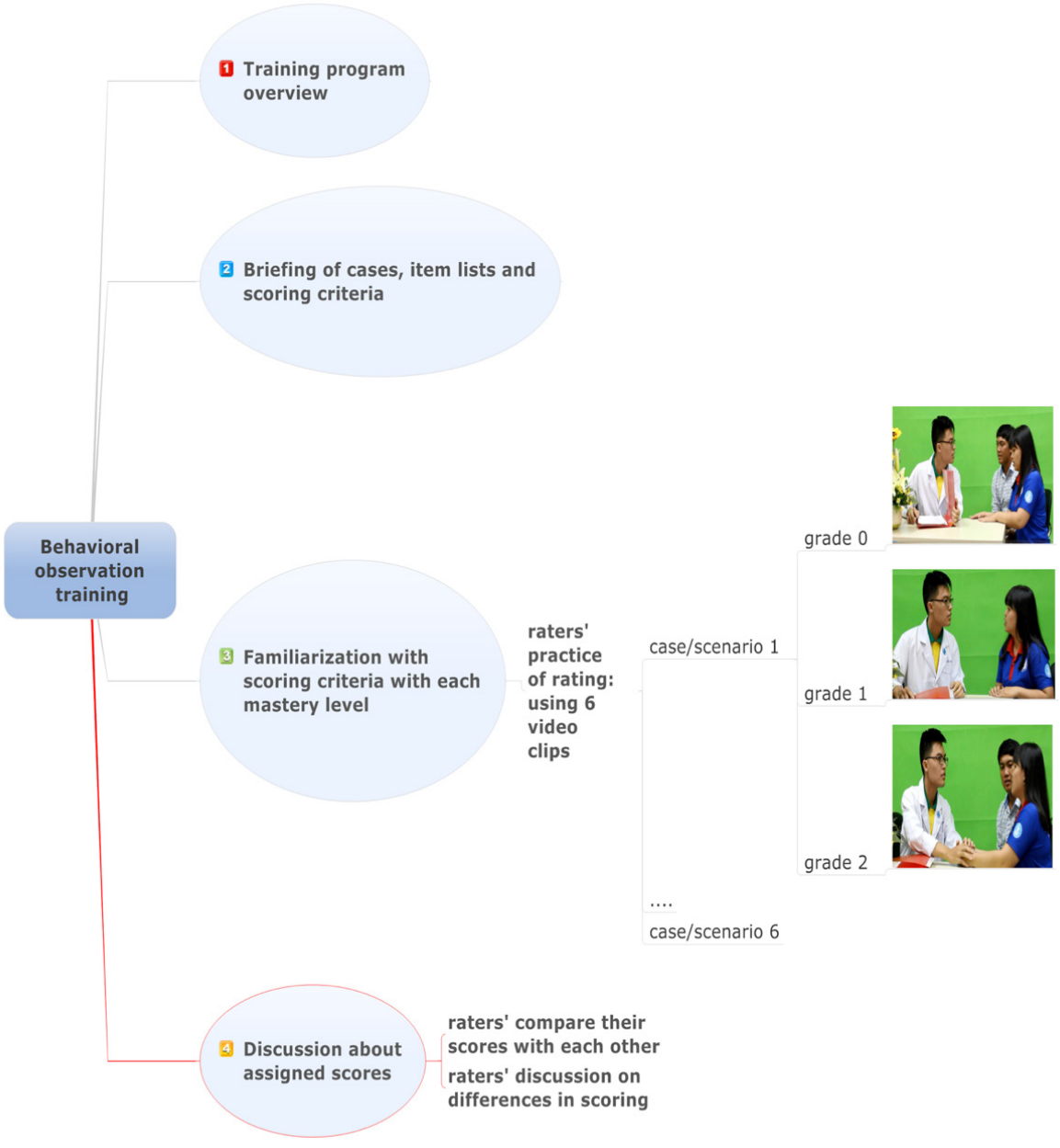

Figure 1: Raters' training process